(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 883 090 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
07.12.2005 Bulletin 2005/49

(51) Int Cl.[7]: **G06T 15/70**

(21) Application number: **98304040.3**

(22) Date of filing: **21.05.1998**

(54) **Method for generating photo-realistic animated characters**

Verfahren zum Erstellen von photo-realistischen beweglichen Figuren

Procédé pour generer des personages animés à réalisme photographique

(84) Designated Contracting States:
DE FR GB IT

(30) Priority: **06.06.1997 US 869531**

(43) Date of publication of application:
09.12.1998 Bulletin 1998/50

(73) Proprietor: **AT&T Corp.**
New York, NY 10013-2412 (US)

(72) Inventors:
- **Cosatto, Eric**
  Highlands, New Jersey 07732 (US)
- **Graf, Hans Peter**
  Lincroft, New Jersey 07738 (US)

(74) Representative: **Hagmann-Smith, Martin P.**
Marks & Clerk,
4220 Nash Court,
Oxford Business Park South
Oxford, Oxfordshire OX4 2RU (GB)

(56) References cited:
WO-A-96/17323 **US-A- 4 841 575**
US-A- 5 367 454

- **ARAD N ET AL: "IMAGE WARPING BY RADIAL BASIS FUNCTIONS: APPLICATION TO FACIAL EXPRESSIONS" CVGIP GRAPHICAL MODELS AND IMAGE PROCESSING, vol. 56, no. 2, 1 March 1994 (1994-03-01), pages 161-172, XP000440387 ISSN: 1077-3169**

EP 0 883 090 B1

Description

FIELD OF THE INVENTION

[0001] This invention relates to the field of animation, and more particularly, to techniques for generating photo-realistic animated faces in computer and communications applications and other high-tech operations.

BACKGROUND OF THE INVENTION

[0002] In a modern technological era characterized by complex computer interfaces and by sophisticated video and communications devices, techniques which promote consumer interest in high-tech equipment have become increasingly prevalent. The computer industry, with an eye toward drawing consumers into the mainstream of its technological innovations, regularly seeks new methods by which consumers can interface with devices that exhibit more user-friendly or humanlike qualities. In addition to increasing the appeal of technology to the average, non-technical consumer, such methods raise the potential for generating a host of new and useful computer and video applications.

[0003] In recent years, the prospect of using animated, talking faces in various high-tech operations has become particularly enticing. Animated techniques can be used for a diverse array of applications. The addition of talking faces to computer-user interfaces, for instance, can enhance the computer's user-friendliness, promote user familiarity with computers, and increase computer entertainment value. Generating animated agents in computer-user interfaces would be useful, among other applications, for performing such tasks as reading E-mail, making announcements, or guiding a user through an application. The use of lifelike, synthesized characters would also be valuable in any application where avatars (visual representations of persons) are used. Such applications could include virtual meeting rooms, multi-user video games, lectures, seminars, security or access devices, scenarios requiring human instructions, and a host of other operations.

[0004] Industry practitioners are also considering model-based coding for video telephony over low data-rate connections. This application would entail generating synthesized faces or characters at the receiver station, whereby the form and structure of the images are governed by parameters originating at the transmitting station.

[0005] Regardless of the particular application, the synthesis of human faces ideally involves the use of lifelike, natural-looking ("photo-realistic") characters. Using photo-realistic characters has many benefits. They have a greater entertainment value over that of simple animated characters. Furthermore, using photo-realistic characters as part of a human interface adds a realistic element to a computer. Consumers otherwise intimidated by computer technology might feel more comfortable using a computer which has a humanlike interface. As another illustration, the use of a photo-realistic character to give an office presentation can create a more favorable impression on the presentation's attendees than if simple animated characters were used - - with simple animation, the characters cannot concurrently speak while eliciting realistic facial and mouth movements typical of a person making speech. Photo-realistic characters can convey meaningful and realistic facial expressions. Simple animation, in contrast, is cartoon-like and unimpressive, particularly in an arena such as a corporate meeting.

[0006] Photo-realistic characters can also be used as an icon in a virtual reality application. Such characters can further be used over such media as the Internet, wherein the bandwidth of the media is otherwise too small to accommodate high-frequency video signals. Using photo-realistic techniques, human characters with realistic movements can be transmitted over the Internet in lieu of video.

[0007] Practitioners have made numerous efforts to synthesize photo-realistic characters. One problem common to most of these methods is their inability to make characters appear sufficiently humanlike. The remaining methods that can, at least in theory, generate more realistic-looking characters require prohibitively large computer memory allocations and processing times to accomplish this goal. The utility of such methods is consequently restricted to high-capacity media.

[0008] Thus, an important but previously unrealized goal of practitioners is to disclose a technique for generating photo-realistic faces which requires a minimal amount of computation for the synthesis of animated sequences. Naturally, the least computation would be required in the case where all parts and their corresponding bitmaps are produced in advance and stored in a library. The synthesis of a face would then involve merely overlaying the parts. Modern graphics processors have become so powerful, however, that warping pictures to generate animated shapes may be performed in real-time. In this event, only the control points need be stored in the library, which substantially reduces the memory required for storing the model.

[0009] The approaches employed by practitioners generally fall into four categories: (1) three-dimensional ("3-D") modeling techniques; (2) warping and morphing techniques; (3) interpolation between views; and (4) flip-book techniques. These approaches are described below.

(1) Three Dimensional Modeling

[0010] Practitioners have developed 3-D models for creating talking heads and faces. Many 3-D modeling techniques use generic mesh models over which pictures of persons are texture-mapped. Generally, the physical structure, such as bone and muscle structure, is designed with great precision and detail to derive the

shape of the face. While 3-D modeling is useful for certain applications, the technique is fraught with disadvantages when deployed in the framework of a character making speech. Such methods require extensive computer processing and consume substantial memory; consequently, they are not suitable for real-time applications on a personal computer. Moreover, facial motions created by standard 3-D models typically seem "robot-like" and do not appear natural.

(2) Warping Or Morphing

[0011] Other techniques for generating animated characters are based on warping or morphing of two-dimensional ("2-D") images of faces. Warping can be generally defined as the intentional distortion of an image in a predefined manner. Warping can be used, *inter alia,* to create expressions on a face by distorting a face with a neutral expression (e.g., to create a frown).

[0012] Morphing is the transformation of one image into another image using interpolation and/or distortion. Unlike warping, which simply distorts an image in a predefined manner, morphing typically uses two sets of fixed parameters comprising an initial image and a target image. Various commercial products utilize warping or morphing techniques, including certain toys which enable children to produce funny, animated sequences of faces.

[0013] One disadvantage of using warping or morphing in isolation is the inability of these techniques to provide realistic, natural-looking facial movements. Another disadvantage is that the morphing process is possible only between predefined pairs of images. Thus this technique is not suitable for generating models that must produce unforeseen motions, such as when previously unknown text is designated to be spoken.

(3) Interpolation Between Reference Views

[0014] Researchers have accordingly sought to resolve shortcomings of existing morphing methods. To accommodate the possibility of extemporaneously generating previously unknown facial motions, researchers have developed techniques to automatically determine morphing parameters. These techniques begin with a set of reference views wherein new views are automatically generated by performing interpolation between the reference views. Researchers have demonstrated sequences of rotating and tilting heads using this interpolation method. However, no sequences of a talking head have been synthesized, in part because the act of speaking generates transient facial features such as grooves and wrinkles that are impossible to duplicate using mere interpolation. Talking also exposes areas not generally seen in the reference frames, such as teeth. Elements like teeth which are not present in the reference frames cannot be generated using interpolation.

(4) Flip-Book Technique

[0015] The flip-book technique, probably the oldest of all animation techniques, involves the process of storing all possible expressions of a face in a memory. Individual expressions are later recalled from the memory to generate a sequence of expressions. The use of this technique has severe practical limitations. For example, making available an adequate number of expressions and mouth shapes requires generating and storing into memory a tremendous number of frames. As such, the flexibility of the flip-book approach is significantly curtailed, since only facial expressions generated ahead of time are available for animation. Thus, among other problems, the flip-book approach is not suitable for use on a personal computer which typically possesses non-trivial limitations with respect to memory size and processing power.

[0016] In sum, the present techniques for synthesizing animated faces possess at least the following disadvantages: (1) the characters portrayed are not sufficiently lifelike or natural-looking, especially when speaking; (2) the methods require a considerable, often infeasible, computer memory allocation and processing time; (3) the characters are frequently incapable of spontaneously uttering words not previously known; (4) the methods lack variety in the possible available facial expressions; (5) many of the methods lack the ability to present various important character features, facial or otherwise; and (6) many of the methods lack coarticulation or conflict resolution techniques.

[0017] Accordingly, it is an object of the invention to disclose a technique for generating lifelike, natural-looking, photo-realistic faces and characters which avoids the complexities of 3-D modeling and which requires less memory and computer-processing power than conventional methods.

[0018] Another object of the invention is to disclose such a method which is capable of generating lifelike facial movements to accommodate concurrent speech, including speech comprising previously unknown text.

[0019] Another object of the invention is to disclose a method which is capable of being executed in a real-time application, including an application on a personal computer.

[0020] Another object of the invention is to disclose a method which provides for a wide variety of possible facial features and expressions.

[0021] US, 4,841,575 discloses an image encoding and synthesis technique. Different images of a mouth are used to generate a moving picture of a speaker.

[0022] WO96/17323 discloses a technique for synthesising animations of a human speaking.

SUMMARY OF THE INVENTION

[0023] The invention provides a method for dealing with synthesized images as set out in the accompanying

claims.

[0024] There is disclosed herein a method for creating photo-realistic, animated faces and characters, such as human and animal faces and characters, which are capable of talking and expressing emotions. The model upon which the method relies is based on one or more pictures of an individual. These pictures are decomposed into a hierarchy of parameterized structures and stored in a model library. For the syntheses of faces, the proper parts are loaded from the model library and overlaid onto a base face to form a whole face with the desired expression.

[0025] To synthesize a wide range of facial expressions, each face part is parameterized using control points stored in memory. These parameterized face parts comprise shapes which cover possible deformations which can be undergone by the face part when executing different actions such as smiling, frowning, articulating a phoneme, etc.

[0026] In a preferred embodiment, the head model is combined with a speech synthesizer from which the model derives the sequence and duration of phonemes to be spoken. The parameters for the face parts are computed from the phoneme sequence.

[0027] Also, in a preferred embodiment, coarticulation effects are computed as a portion of the computations for one or more face parts. Conflicts are preferably resolved between the requirements for expressing emotion and for saying a phoneme. A version of each face part matching the desired values is thereupon generated from the library for the synthesis of the whole face.

BRIEF DESCRIPTION OF THE DRAWINGS

[0028]

Fig. 1 shows the decomposition of a face into parameterized parts according to the method of the present invention.

FIGS. 2a and 2b, collectively FIG. 2, shows a method of parameterizing face parts in accordance with a preferred embodiment of the invention.

FIG. 3 shows three visemes generated from photographs, in accordance with a preferred embodiment of the invention.

FIG. 4 shows three visemes generated by warping, in accordance with a preferred embodiment of the invention.

FIG. 5 shows a process of transitioning from a neutral expression to a smile in accordance with a preferred embodiment of the invention.

FIG. 6 shows an exemplary flowchart illustrating a synthesis of the animation sequence in accordance with a preferred embodiment of the invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0029] The method according to the present invention is well-suited, among other operations, for virtually any application involving reading text or giving a speech to a user or other person. While the invention can be fully understood from the following explanation of its use in the context of human faces and characters, the invention is intended to be equally applicable to the use of animal faces and characters. In a preferred embodiment, the technique according to the invention comprises four groups of elements:

(1) base face: a picture of a face which is used as a background for the animated sequence. Typically, although not necessarily, the base face contains a neutral expression and a closed mouth.

(2) parameterized facial parts ("PFPs"): individual facial parts which are overlaid onto the base face during the course of the animation sequence. Preferably, PFPs comprise the mouth, eyes, nose, cheeks, chin, forehead, teeth, tongue, and mouth cavity.

(3) parameterized facial parts with hair ("PFPHs"): individual facial parts with hair which are overlaid onto the base face during the course of the animation sequence. Preferably, PFPHs comprise the hair, eyebrows, moustache, beard, and side burns.

(4) secondary effects: additional features which may be used during the animation sequence. These may include, for example, grooves, wrinkles, and highlights.

These four groups of elements are described in greater detail below in the context of various preferred embodiments of the invention.

(1) Base Face

[0030] The base face is a picture of the whole face onto which parts are overlaid for the animation sequence. The use of a base face in an animation sequence can save computer processing time and reduce memory requirements. This is because for many expressions, only a particular section of the base face need be changed; all other sections can remain undisturbed.

[0031] If a facial image contains pronounced wrinkles and grooves, these are preferably removed before using the image as a base face. Such a removal procedure may be necessary to avoid interference with wrinkles and grooves added during the course of the animation sequence.

(2) Parameterized Facial Parts ("PFP")

[0032] The PFPs are the individual elements that are overlaid onto the base face to comprise the image of the whole face for an animated frame. The features of PFPs are defined by one or more parameters. Illustrations of parameters include shape and depth. The parameters of each PFP are stored as control points in a memory. In a preferred embodiment, a PFP is represented by one or more rectangular bitmaps, with each bitmap comprising a set of parameters and a mask. In this embodiment, the control points are positioned on the bitmap. A bitmap may be used, for example, to represent a PFP with a distinct shape. A related basemap may contain the actual pixels of the picture to be displayed.

[0033] The mask defines the transparency of each pixel comprising the image. Thus, the mask may describe which part inside the bounding box is visible and how the transparency gradually increases at the borders to merge the PFP with the background.

[0034] Examples of PFPs, and how they relate to a whole face are shown in FIG. 1. (The procedure by which PFPs are parameterized is explained in greater detail in FIG. 3.) FIG. 1 shows a face which has been decomposed into PFPs. In this illustration, the picture 100 of the whole face is divided into the following parts: forehead 110, eyes 120, nose 130, mouth 140, hair 150, cheeks 160, teeth 170, and chin 180. The forehead 110 has been retouched to remove hair. All other parts are left in their original state.

[0035] In a preferred embodiment, parameters stored with each PFP include at least its width and height, and additional parameters defining how to place the part onto the image. As an illustration, for a mouth, points corresponding to the edges of the upper and lower teeth may be used to define the mouth's placement (see FIG. 2).

[0036] Other parameters may be used to define the outline of a face part and the outline of other prominent features. The inventors presently prefer outlines encoded with splines. A spline generally uses a piecewise function approximation of the outline, such as a third-order polynomial. This polynomial function describes the outline between two control points. A second polynomial function describes the outline between the next two points, and so on. For the mouth, the contours of the lips may be encoded in this manner. For the eyes, the outline of the shape, the contour of the eye opening, and the upper edge of the eye lid may be encoded.

[0037] Another parameter which can be useful is the depth of the PFP, i.e., the position of the PFP on the axis perpendicular to the bitmap relative to the other PFPs. Using this parameter, the PFP can take the 3-D structure of the face into account. Thus the present invention can incorporate various assets of 3-D modeling without the associated liabilities of increased memory allocation and processing time.

[0038] For each PFP, several bitmaps may be stored, with each bitmap comprising a representation of the PFP in a different shape. The different shapes of the PFP can, in turn, be obtained by decomposing separate pictures of the face in a process called photographic extraction. Each of these pictures may have a different expression, and the individual PFPs unique to that expression can be extracted and stored in memory as one or more bitmaps. Alternatively, or in addition to this method, the shapes of the PFPs may be produced by other methods of decomposition: warping and/or morphing from key shapes. The preferred methods for decomposing pictures to obtain PFPs are explained more fully below.

[0039] An example of the parameterization of a face part is shown in FIG. 2. The image 200 comprising FIG. 2a illustrates the parameters of the mouth used to place it and to define its appearance. These parameters are inside height (H-m1), outside height (H-m2), and width (W-m). The point c comprises the lower edge of the upper teeth, and the point d comprises the upper edge of the lower teeth. The points c and d are used to position the mouth onto the image. The placement of the mouth is performed relative to the points a (center between eyes) and b (center between nostrils). In this illustration, all distances are normalized to the eye separation (W-e). Eye-nose (H-en) and eye-mouth (H-em) distances are used to estimate the head tilt. FIG. 2b shows the lip contours 210, which comprise the curves describing the inner contours of the lips (ic), the outer contour (oc), and the shape, or border, of the part. Marked on the lip contours 210 are control points 220 defining the deformation in a warping or morphing operation. As in this example, warping or morphing may be used to manipulate individual PFPs.

[0040] As described above, PFPs may be generated by decomposing pictures of a face. One embodiment of this decomposition method is called photographic extraction. Photographed parts provide a good base for generating animated sequences which closely resemble the photographed person. To generate the shapes, a person may be asked to speak a phoneme or to express an emotion. From such pictures the shapes are simply cut out, and scanned or otherwise placed into memory.

[0041] Two additional considerations are important to the process of photographic extraction. First, the total number of frames required to cover all possible shapes for a given animation sequence may be large. Second, while the quality of the individual frames synthesized using this photographic extraction method tends to be high, the practitioner must ensure that the individual features look natural in the context of the whole facial expression. This latter consideration is related to the action of the person being photographed. Specifically, when a person purposefully expresses an emotion or speaks a phoneme in isolation, he or she tends to exaggerate the articulation. Thus a photographed shape is often not immediately appropriate for animation.

[0042] For these reasons, pictures generated via photographic extraction usually are insufficient to create a complete set of shapes for a particular application (i.e., a set where all possible shapes to be encountered in an animation sequence have been produced). These pictures are nevertheless an important part of the process for generating a complete set of shapes. Additional shapes may be produced by using morphing, warping, and/or interpolation techniques. Using morphing and/or interpolation, all shapes which are intermediate to a neutral expression and an exaggerated one may be generated. Among these intermediate images, several may be suitable for animation.

[0043] In another embodiment, warping alone may be used as the decomposition method to create the library of PFPs. Here, animated frames are produced by generating all necessary shapes from a single shape. Generating facial parts by warping entails (1) referencing the control points in memory which define an original facial shape; (2) adjusting those control points into new control points which define the new deformed shape; and (3) recording the new control points in memory for the new shape. Not surprisingly, this method generally requires a detailed understanding of how a facial part deforms.

[0044] Once a library of individual shapes are created by warping, they may advantageously be used in any number of images of photo-realistic persons since facial parts are individually parameterized. By dictating the new control points, the splines describing the feature shapes may be easily adapted to take into account the characteristics of a different person (e.g., width of the face, thickness of the lips, etc.).

[0045] While the final photo-realistic person created using facial parts derived by warping may not be exactly like how the person really looks, the warping suffices to create a reasonably realistic representation. Moreover, warping is fast and convenient, since a single picture of a person may be used to generate all of the animated frames.

[0046] The process of generating images of the mouth is probably the most complex of all facial features. This is particularly true where the photo-realistic person is speaking. Indeed, the mouth shows the widest variations of all PFPs. In a talking face, it is also the feature to which the observer is most attentive. Human beings are sensitive to slight irregularities in the shape or motion of the mouth. Hence, the practitioner should ordinarily pay special attention to the animation of the mouth.

[0047] A mouth shape articulating a phoneme is often referred to as a viseme. While over fifty spoken visemes are distinguished in the English language, most researchers consider between ten and twenty different visemes to be sufficient to use in an animated sequence. The number of visemes, of course, will vary depending on the application. In a preferred embodiment of the invention, twelve principle visemes are employed, namely: a, e, ee, o, u, f, k, l, m, t, w, mouth-closed. All other possible phonemes are mapped onto this set of twelve.

[0048] FIGS. 3 and 4 show examples of visemes, generated using the two techniques described above. The visemes in FIG. 3 are cut out from two separate pictures. In FIG. 3, the mouth area 300 was cut out from a photograph of the person speaking the phoneme "u". Similarly, the areas 310 and 320 were cut out from photographs of the person speaking the phonemes "m" and "a", respectively. The bottom row shows the three visemes overlaid onto the base face. The resulting face was then placed onto a background to form frames 330, 340, and 350. The lady in FIG. 3 was speaking the phonemes in isolation; as such, they appear strongly articulated.

[0049] The visemes in FIG. 4 are generated by warping from a single picture. Thus, a picture of the person was photographed, and from this picture all visemes and other expressions were generated by warping. Visemes 410, 420, and 430 were generated using the phonemes "u", "m", and "a", respectively. The bottom row shows the three frames 440, 450, and 460 with visemes 410, 420, 430 overlaid onto the base face together with variations of the eyes and eyebrows.

[0050] Judging these individual frames, most people would consider the appearance of the faces in FIG. 3 more natural than those in FIG. 4. Yet, when the visemes of FIG. 3 are used for animation, the result is a jerky, exaggerated motion of the mouth that looks unnatural. The much less pronounced articulation produced with the visemes of FIG. 4 is perceived as more resembling a real person. This observation highlights the significance of designing the motion rather than solely concentrating on the appearance of the individual frames. For this reason, it is preferable to generate two or three versions of each viseme, each version representing a different strength of articulation.

[0051] In practice, it is possible to achieve truly natural-appearing mouth motion only when coarticulation effects are taken into account. Coarticulation means that the appearance of a mouth shape depends not only on the phoneme produced at the moment, but also on the phonemes preceding and succeeding that phoneme. For instance, when an individual articulates the phrase "boo", the mouth shape for "b" reflects the intention of the individual to say "oo". In short, accounting for articulation generates a smoother mouth animation, which, in turn, avoids unnatural exaggeration of lip motion during the animation sequence.

[0052] Accordingly, a preferred embodiment of the invention uses coarticulation. The preferred coarticulation method involves the assignment of a mathematical time constant to the parameters of the mouth. Using this time constant, the present shape of the mouth can be made to influence the manner and extent to which the mouth can deform in the proceeding time interval.

(3) Parameterized Facial Parts With Hair ("PFPH")

[0053] The PFPH is a group that comprises all of the

parts of the face covered with hair. In a preferred embodiment of the invention, the PFPHs are grouped separately from the PFPs because their deformations are typically processed differently. Standard morphing and warping techniques tend to smear out the characteristic textures of hair, or they deform the textures in ways which look artificial. Filling contours with copied textures usually yields better results for moustaches and beards.

[0054] In some embodiments, hair is animated only in a limited manner. If the internal texture of the hair is not recognizable, hair can simply be treated like a smooth surface. For images having very low resolution, this crude modeling may be appropriate. However, if individual bangs of hair or even individual hairs are visible, a more sophisticated approach is preferred. For example, motion may be added to whole bangs of hair by rotating and skewing parts of the hair. In actuality, motion of the hair has a limited significance for a talking head. Therefore, only some random motions need be added to the hair.

[0055] The most explicit and deterministic motion of a hairy part in a talking face is that of a moustache. A moustache can be deformed by skewing and by cut-and-paste operations. For example, when the mouth changes from a neutral expression to saying an 'o', the outline of the moustache may be estimated from the shape of the upper lip. Thereupon, the original moustache is bent to follow the upper contour using local skew operations. To the extent that parts of the contour are left blank, neighboring sections may be copied and filled into blank sections.

(4) Secondary effects

[0056] The generation of realistic-looking images involves additional intricacies. Such intricacies include wrinkles, grooves, and highlights. Extreme deformations, such as a broad smile, are difficult to generate without the addition of grooves. FIG. 5 illustrates the effect of adding grooves to a smile. Rather than generating PFPs with grooves, the grooves are overlaid onto the PFP.

[0057] FIG. 5 shows a frame 510 which has a background face and a neutral expression. Frame 520 has the same background face but instead has eyes 515 and mouth 518 overlaid on the background frame. The result of Frame 520 is an unnatural expression. Frame 530 is the same frame as frame 520 except that frame 530 has grooves 535 overlaid onto its base face. The smile of the character in frame 530 looks more natural as a result of the grooves.

[0058] Grooves and wrinkles are preferably categorized as a distinct group of the head model because both their syntheses and deformations are treated differently than those of the PFPs. The warping and morphing used to generate the different shapes of the PFP abnormally distort the grooves. Thus, in a preferred embodiment, grooves and wrinkles are instead represented by splines. Splines define the position of the grooves. Adding grooves to a bitmap may be achieved by modulating the color of pixels with a luminance factor. The groove defines the extent of this modulation as well as the gradients in the direction perpendicular to the groove direction.

[0059] For the synthesis of a naturally-looking talking head, the motions of all facial parts and of the head must be scrupulously planned. Conversational signals comprise subtle movements of facial parts of the head that punctuate, emphasize, or otherwise regulate speech. For example, a rising eyebrow can be used to accentuate a vowel or to indicate a question. Eye blinks also occur frequently and are usually synchronized with speech flow. Slight head movements also generally accompany speech. When such motions stop, it often means that the speaker has finished and is expecting the listener to take some action. The emotional state of the speaker is also reflected by changes in face parts appearance. For example, eyebrows raised and drawn together may indicate tension or fear.

[0060] One illustration of the synthesis process is shown in FIG. 6. In response to ASCII input comprising the desired words to be spoken (oval 600), a text-to-speech synthesizer (box 610) produces a sequence of phonemes 620, their duration and stress. Each phoneme is mapped to a mouth viseme (box 630). In some embodiments, this mapping can comprise a simple lookup operation such as a table in memory. Once a viseme has been selected, the parameters of the viseme 640 are available for that viseme. These parameters may include, for example, the width and height of the mouth.

[0061] The parameters of the viseme may thereupon be entered into a coarticulation module to account for coarticulation effects (box 660). The coarticulation module can also take into account information relating to the desired facial expression (oval 650). As an illustration, if a smile is requested and the viseme calls for a closed mouth, the coarticulation module will increase the mouth width. The output of the coarticulation module is a new set of mouth parameters 670. These modified parameters are next used in the search of the PFP library for the shape with the closest match. The closest-matched PFP is selected (box 680).

[0062] The other PFPs are preferably selected using the same process. For those PFPs affected by mouth motion, phoneme information as well as facial expression information is considered. For the eyes and everything above the eyes, only facial expressions need be taken into account since mouth motion does not usually affect these parts.

[0063] After the proper visemes have been selected, they are ready to be blended into the base face for the ultimate generation of a frame of the animated sequence (box 690). The frame is synchronized with the corresponding speech. Where the visemes are generated using warping techniques, they will no longer

seamlessly blend into the base head. For that reason, the practitioner must perform careful feathering of the alpha-blending mask. In a preferred embodiment, blending is performed first with the parts lying deepest, such as eyes and teeth. Next, the parts lying above are added, and finally, the hair and wrinkles.

[0064] Once the head contains all face parts, it can be overlaid onto the background image. Motion of the whole head may then be added. Motion vectors are computed in a semi-random manner, e.g., speech regulates a random blend of predefined motions. A model according to a preferred embodiment includes translating, rotating, and tilting of the head. Rotation of the head around the axis perpendicular to the image plane is readily accomplished. Small rotations around the other two axes may be approximated by simple and fast image warping techniques. The dynamics of the motions must be carefully designed -- exaggerated motions appear jerky and unnatural.

[0065] The frame is then output to a file or a screen (box 695). In some embodiments, these pixel operations can be performed either into a frame buffer in computer memory or directly into the frame buffer of the screen. At the time of filing this application, the inventors had implemented at least two versions of the invention. One version used the Microsoft AVI API and generated AVI files from an ASCII text input. The other version outputted animation directly onto the screen using the OpenGL graphics library.

[0066] Also, at the time of filing this application, the inventors were running an experimental implementation of the head model on a personal computer (PentiumPro 150 MHZ). In this implementation, the synthesizing rate was approximately one frame every 100 milliseconds. The size of the PFP library was 500 Kilobytes. The PFP library contained all visemes for talking and all PFPs for frowning and smiling, happy and neutral expressions. The size of the frames was 360 x 240 pixels. It should be understood that parameters will vary widely depending on the particular embodiment of the invention. The inventors' experiment was designed for maximum flexibility rather than for ultimate speed or maximum compactness. Further, while the present invention contemplates a wide variety in numerical ranges depending on the state of technology and on the particular application, etc., the inventors estimate that an optimization for speed would permit a synthesis at nearly 30 frames per second and that the library size can be reduced by more than a factor of two. Without using the method according to this invention, the inventors estimate that a library of faces covering the same range of expressions would be more than an order of magnitude larger.

[0067] It will be understood that the foregoing is merely illustrative of the principles of the invention, and that various modifications and variations can be made by those skilled in the art without departing from the scope of the invention as defined by the claims.

Claims

1. A method for synthesizing an animation associated with a phoneme sequence, comprising:

mapping each phoneme in a phoneme sequence to a viseme (630);
selecting one or more parameterized animation parts from a library of parameterized animation parts based on mapped viseme parameters (640); and
overlaying the one or more parameterized animation parts on a base animation to synthesize an animation (690),

characterized in that the library of parameterized animation parts is populated by decomposing at least one image into a hierarchy of parameterized animation parts and **in that** the parameters associated with each parameterized animation part enable deformation of the animation part when synthesizing the animation.

2. The method of claim 1, further comprising:

modifying the viseme parameters prior to selecting the one or more parameterized animation parts, wherein the selection of the parameterized animation parts is based on the modified viseme parameters (680).

3. The method of claim 2, wherein the step of modifying the viseme parameters is further **characterized by** calculating coarticulation effects (660).

4. The method of claim 3, wherein the step of modifying the viseme parameters is further performed using information obtained about a desired expression (650).

5. The method of claim 1, wherein the parameters corresponding to the mapped viseme are mouth width and mouth height.

6. The method of claim 3, wherein a coarticulation module performs the step of calculating coarticulation effects (660) and receiving information related to the desired expression (650) and based on the calculated and received data, the coarticulation module modifies the mapped viseme parameters for selecting the one or more parameterized animation parts.

7. The method of claim 2, wherein the modified viseme parameters are used to generate deformed parameterized animation parts for storage in a library.

8. The method of claim 7, wherein the deformed pa-

rameterized animation parts are generated from pictures of a person.

9. The method of claim 2, wherein the modified viseme parameters are used to generate deformed parameterized animation parts dynamically when synthesizing the animation.

10. The method of claim 9, wherein dynamically synthesizing the animation using deformed parameterized animation parts further comprises deforming shapes associated with the parameterized animation parts.

11. The method of claim 1, wherein the animation is a facial animation.

12. The method of claim 1, wherein the library of parameterized animation parts comprises a decomposition of pictures of an individual or object into a hierarchy of parameterized animation parts.

13. The method of claim 12, wherein decomposing the pictures further comprises one or more processes from a set that includes photographic extraction, warping, morphing and interpolation.

14. The method of claim 13, wherein the photographic extraction step further comprises interpolation between reference views.

15. The method of claim 11, wherein the facial animation comprises a mouth, eyes, a nose, cheeks, a chin, a forehead, teeth, a tongue, and a mouth cavity.

16. The method of claim 1, wherein the phoneme sequence is derived from text to be spoken by the animation.

Patentansprüche

1. Verfahren zum Synthetisieren einer mit einer Phonemsequenz verknüpften Animation, mit den Schritten:

Zuordnen jedes Phonems in einer Phonemsequenz zu einem Visem (630);
Wählen eines oder mehrerer parametrierter Animationsteile aus einer Bibliothek von parametrierten Animationsteilen auf der Grundlage von zugeordneten Visem-Parametern (640); und
Anordnen des einen oder der mehreren parametrierten Animationsteile auf einer Basisanimation, um eine Animation (690) zu synthetisieren,

dadurch gekennzeichnet, daß die Bibliothek von parametrierten Animationsteilen **dadurch** bestückt wird, daß zumindest ein Bild in eine Hierarchie von parametrierten Animationsteilen zerlegt wird, und daß die mit jedem parametrierten Animationsteil verknüpften Parameter eine Verformung des Animationsteils ermöglichen, wenn die Animation synthetisiert wird.

2. Verfahren nach Anspruch 1, ferner mit dem Schritt:

Modifizieren der Visem-Parameter vor dem Wählen des einen oder der mehreren parametrierten Animationsteile, wobei die Wahl der parametrierten Animationsteile auf den modifizierten Visem-Parametern (680) beruht.

3. Verfahren nach Anspruch 2, wobei der Schritt des Modifizierens der Visem-Parameter ferner **gekennzeichnet ist durch** Berechnen von Koartikulationseffekten (660).

4. Verfahren nach Anspruch 3, wobei der Schritt des Modifizierens der Visem-Parameter ferner unter Verwendung von Information durchgeführt wird, die über einen gewünschten Ausdruck (650) gewonnen wird.

5. Verfahren nach Anspruch 1, wobei die Parameter, die dem zugeordneten Visem entsprechen, Mundbreite und Mundhöhe sind.

6. Verfahren nach Anspruch 3, wobei ein Koartikulationsmodul den Schritt des Berechnens von Koartikulationseffekten (660) und des Empfangens von Information, die sich auf den gewünschten Ausdruck (650) bezieht durchführt und auf Basis der berechneten und empfangenen Daten das Koartikulationsmodul die zugeordneten Visem-Parameter zum Wählen des einen oder der mehreren parametrierten Animationsteile modifiziert.

7. Verfahren nach Anspruch 2, wobei die modifizierten Visem-Parameter verwendet werden, um verformte parametrierte Animationsteile zur Speicherung in einer Bibliothek zu erzeugen.

8. Verfahren nach Anspruch 7, wobei die verformten parametrierten Animationsteile aus Bildern einer Person erzeugt werden.

9. Verfahren nach Anspruch 2, wobei die verformten Visem-Parameter verwendet werden, um verformte parametrierte Animationsteile dynamisch zu erzeugen, wenn die Animation synthetisiert wird.

10. Verfahren nach Anspruch 9, wobei eine dynamische Synthetisierung der Animation unter Verwen-

dung verformter parametrierter Animationsteile ferner den Schritt umfaßt: Verformen von Formen, die mit den parametrierten Animationsteilen verknüpft sind.

11. Verfahren nach Anspruch 1, wobei die Animation eine Gesichtsanimation ist.

12. Verfahren nach Anspruch 1, wobei die Bibliothek von parametrierten Animationsteilen eine Zerlegung von Bildern eines Individuums oder Objekts in eine Hierarchie von parametrierten Animationsteilen umfaßt.

13. Verfahren nach Anspruch 12, wobei die Zerlegung der Bilder ferner einen oder mehrere Prozesse aus einer Menge von Prozessen umfaßt, die fotografische Extraktion, Verzerrung, Morphing und Interpolation umfaßt.

14. Verfahren nach Anspruch 13, wobei der fotografische Extraktionsschritt ferner eine Interpolation zwischen Referenzansichten umfaßt.

15. Verfahren nach Anspruch 11, wobei die Gesichtsanimation einen Mund, Augen, eine Nase, Wangen, ein Kinn, eine Stirn, Zähne, eine Zunge und eine Mundhöhle umfaßt.

16. Verfahren nach Anspruch 1, wobei die Phonemsequenz von dem von der Animation zu sprechenden Text abgeleitet wird.

Revendications

1. Procédé pour synthétiser une animation associée à une séquence de phonèmes, comprenant:

la cartographie de chaque phonème dans une séquence de phonèmes selon un visème (630);
la sélection d'une ou de plusieurs parties d'animation paramétrées à partir d'une bibliothèque de parties d'animation paramétrées sur la base de paramètres de visème cartographiés (640); et
la mise en chevauchement des unes ou plusieurs parties d'animation paramétrées sur une animation de base afin de synthétiser une animation (690),

caractérisé en ce que la bibliothèque de parties d'animation paramétrées est peuplée en décomposant au moins une image selon une hiérarchie de parties d'animation paramétrées et **en ce que** les paramètres qui sont associés à chaque partie d'animation paramétrée permettent la déformation de la partie d'animation lors de la synthèse de l'animation.

2. Procédé selon la revendication 1, comprenant en outre:

la modification des paramètres de visème avant la sélection des une ou plusieurs parties d'animation paramétrées, où la sélection des parties d'animation paramétrées est basée sur les paramètres de visème modifiés (680).

3. Procédé selon la revendication 2, dans lequel l'étape de modification des paramètres de visème est en outre **caractérisée par** le calcul d'effets de coarticulation (660).

4. Procédé selon la revendication 3, dans lequel l'étape de modification des paramètres de visème est en outre réalisée en utilisant une information qui est obtenue en ce qui concerne une expression souhaitée (650).

5. Procédé selon la revendication 1, dans lequel les paramètres correspondant au visème cartographié sont la largeur de bouche et la hauteur de bouche.

6. Procédé selon la revendication 3, dans lequel un module de coarticulation réalise l'étape de calcul d'effets de coarticulation (660) et de réception d'une information rapportée à l'expression souhaitée (650), et sur la base des données calculées et reçues, le module de coarticulation modifie les paramètres de visème cartographiés pour sélectionner les unes ou plusieurs parties d'animation paramétrées.

7. Procédé selon la revendication 2, dans lequel les paramètres de visème modifiés sont utilisés pour générer des parties d'animation paramétrées et déformées pour un stockage dans une bibliothèque.

8. Procédé selon la revendication 7, dans lequel les parties d'animation paramétrées et déformées sont générées à partir d'images d'une personne.

9. Procédé selon la revendication 2, dans lequel les paramètres de visème modifiés sont utilisés pour générer des parties d'animation paramétrées et déformées de manière dynamique lors de la synthèse de l'animation.

10. Procédé selon la revendication 9, dans lequel une synthèse dynamique de l'animation en utilisant des parties d'animation paramétrées et déformées comprend en outre la déformation de formes qui sont associées aux parties d'animation paramétrées.

11. Procédé selon la revendication 1, dans lequel l'animation est une animation de visage.

12. Procédé selon la revendication 1, dans lequel la bibliothèque de parties d'animation paramétrées comprend une décomposition d'images d'un individu ou d'un objet selon une hiérarchie de parties d'animation paramétrées.

13. Procédé selon la revendication 12, dans lequel la décomposition des images comprend en outre un ou plusieurs processus pris parmi un jeu qui inclut une extraction photographique, un gauchissement, un morphage et une interpolation.

14. Procédé selon la revendication 13, dans lequel l'étape d'extraction photographique comprend en outre une interpolation entre des vues de référence.

15. Procédé selon la revendication 11, dans lequel l'animation de visage comprend une bouche, des yeux, un nez, des joues, un menton, un front, des dents, une langue et une cavité buccale.

16. Procédé selon la revendication 1, dans lequel la séquence de phonème est dérivée à partir d'un texte destiné à être parlé par l'animation.

FIG. 1

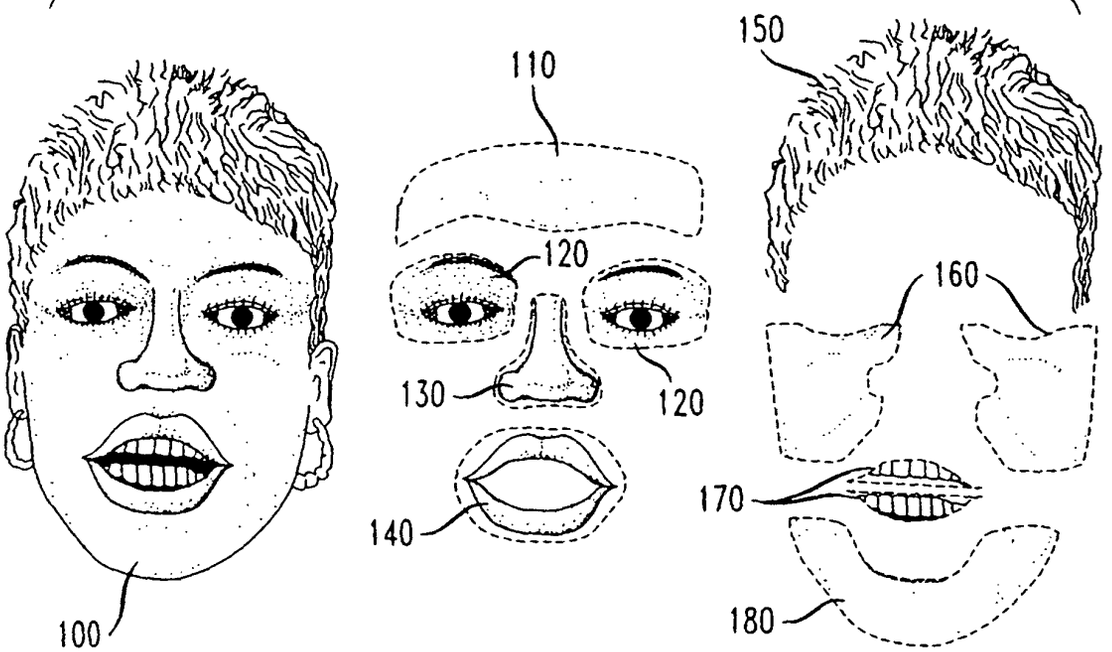


FIG. 2A

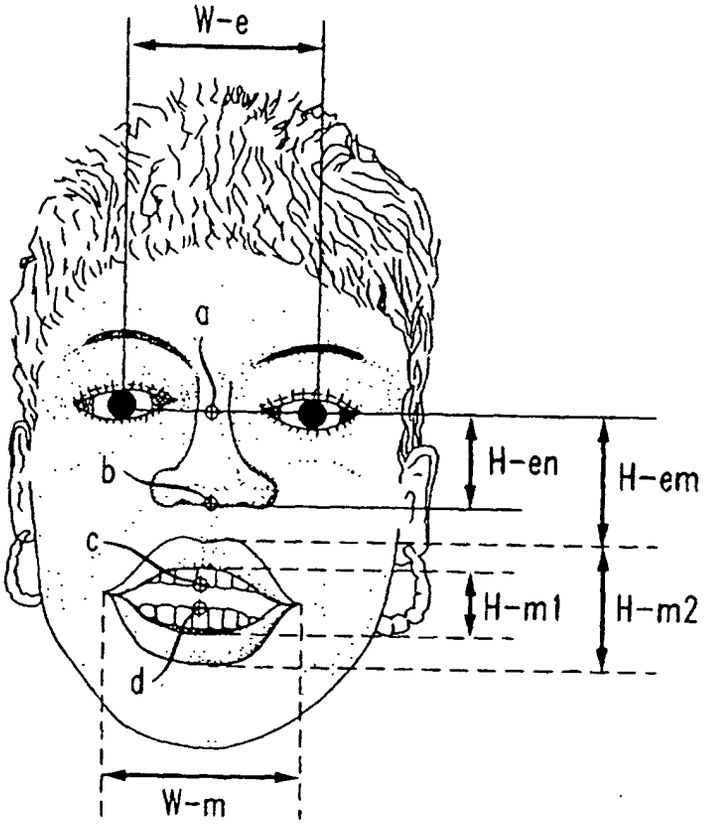


FIG. 2B

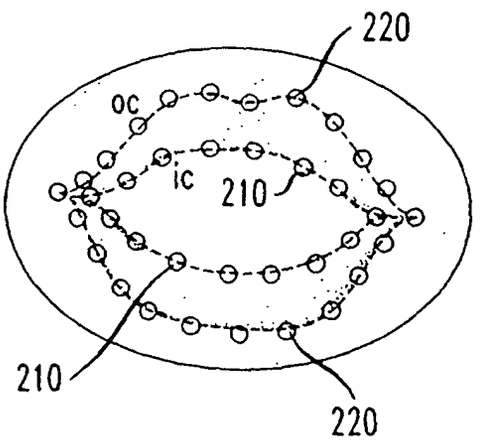


FIG. 3

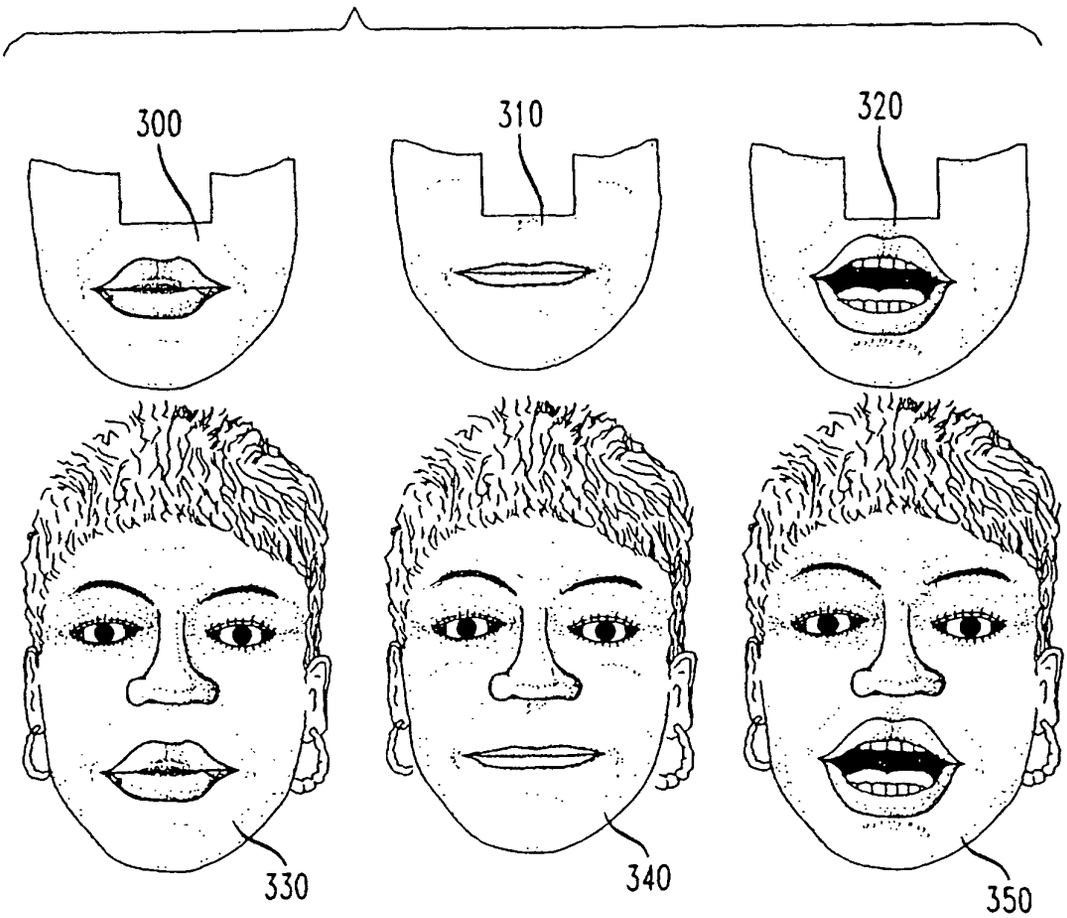


FIG. 4

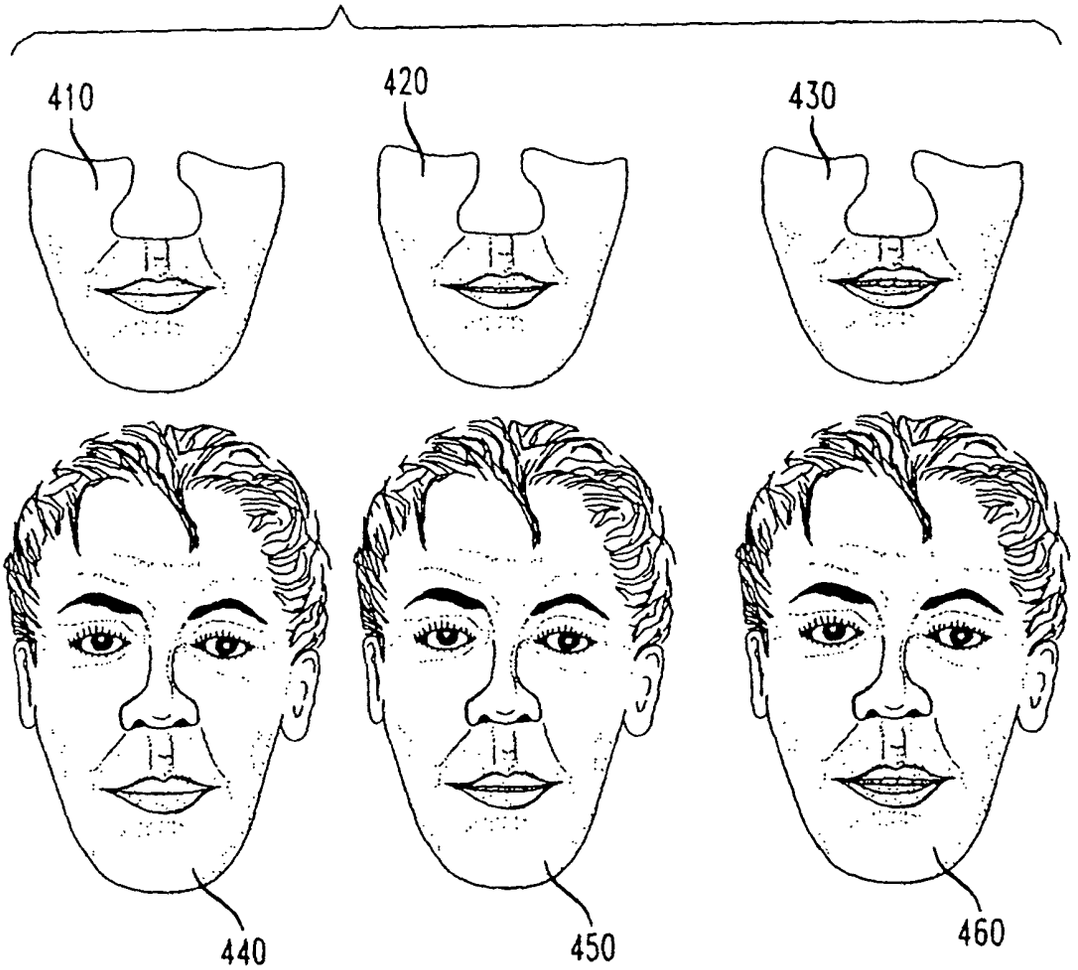


FIG. 5

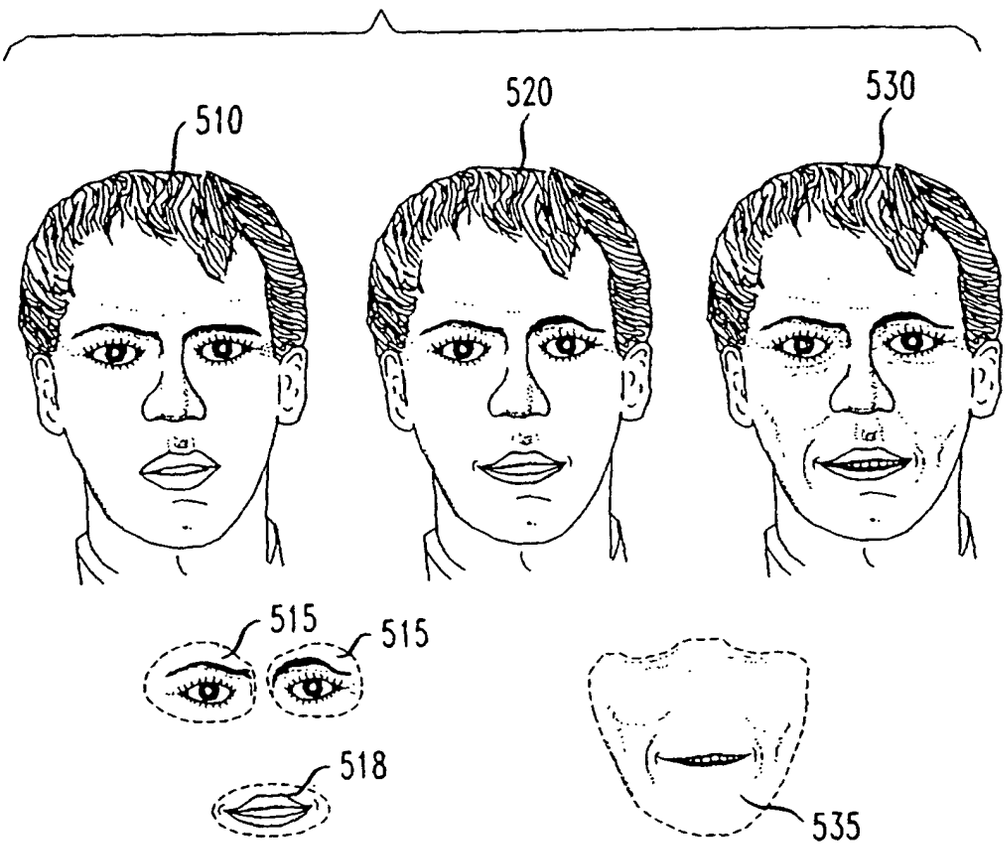
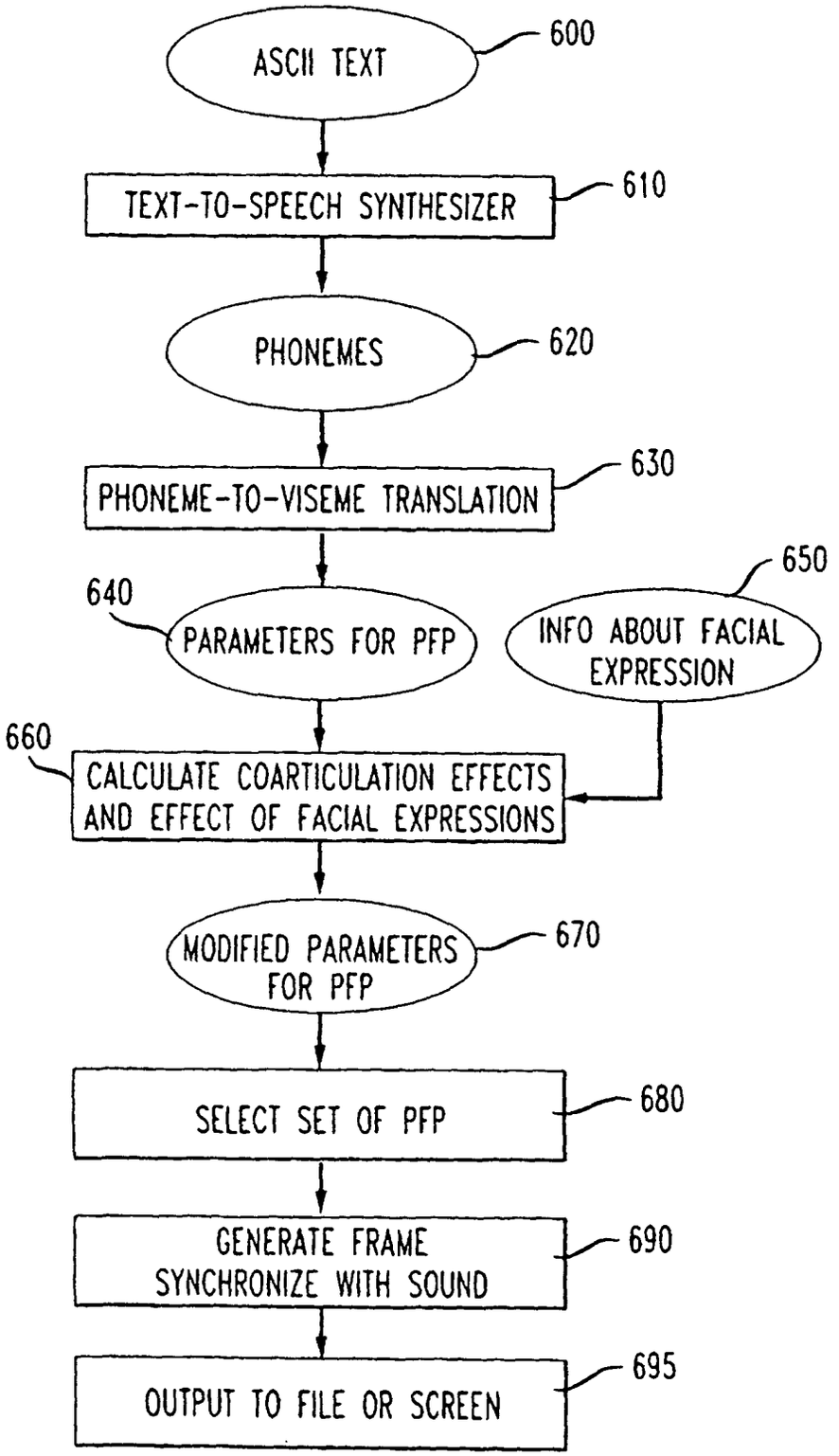


FIG. 6


ASCII TEXT
600
TEXT-TO-SPEECH SYNTHESIZER
610
PHONEMES
620
PHONEME-TO-VISEME TRANSLATION
630
640
PARAMETERS FOR PFP
650
INFO ABOUT FACIAL EXPRESSION
660
CALCULATE COARTICULATION EFFECTS AND EFFECT OF FACIAL EXPRESSIONS
MODIFIED PARAMETERS FOR PFP
670
SELECT SET OF PFP
680
GENERATE FRAME SYNCHRONIZE WITH SOUND
690
OUTPUT TO FILE OR SCREEN
695